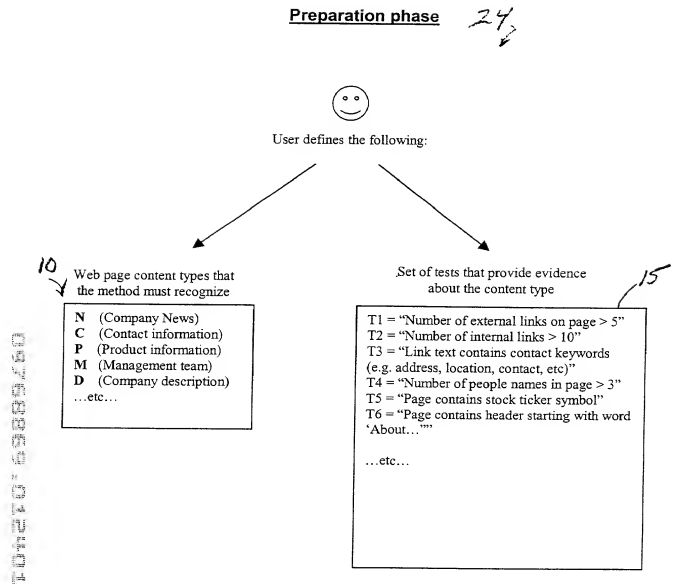
Preparation phase
24
User defines the following:
10
Web page content types that the method must recognize
N (Company News)
C (Contact information)
P (Product information)
M (Management team)
D (Company description)
…etc…
Set of tests that provide evidence about the content type
15
T1 = "Number of external links on page > 5"
T2 = "Number of internal links > 10"
T3 = "Link text contains contact keywords (e.g. address, location, contact, etc)"
T4 = "Number of people names in page > 3"
T5 = "Page contains stock ticker symbol"
T6 = "Page contains header starting with word 'About…""
…etc…

Fig. 1

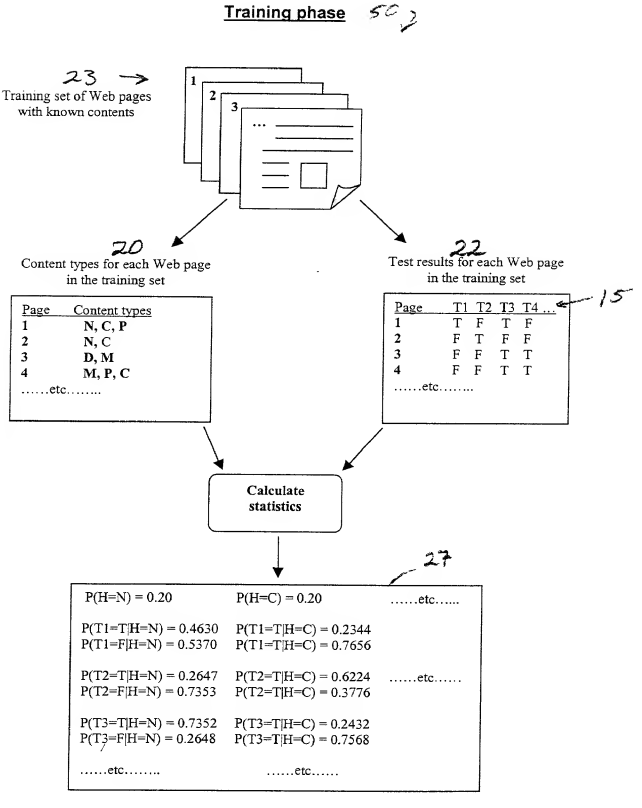
Training phase
50
23
Training set of Web pages with known contents
1
2
3
...
20
Content types for each Web page in the training set
22
Test results for each Web page in the training set
15
Page Content types
1 N, C, P
2 N, C
3 D, M
4 M, P, C
.....etc.....
Page T1 T2 T3 T4 ...
1 T F T F
2 F T F F
3 F F T T
4 F F T T
.....etc.....
Calculate statistics
27
P(H=N) = 0.20
P(H=C) = 0.20
.....etc.....
P(T1=T|H=N) = 0.4630
P(T1=F|H=N) = 0.5370
P(T1=T|H=C) = 0.2344
P(T1=T|H=C) = 0.7656
P(T2=T|H=N) = 0.2647
P(T2=F|H=N) = 0.7353
P(T2=T|H=C) = 0.6224
P(T2=T|H=C) = 0.3776
.....etc.....
P(T3=T|H=N) = 0.7352
P(T3=F|H=N) = 0.2648
P(T3=T|H=C) = 0.2432
P(T3=T|H=C) = 0.7568
.....etc.....
.....etc.....

Fig. 2

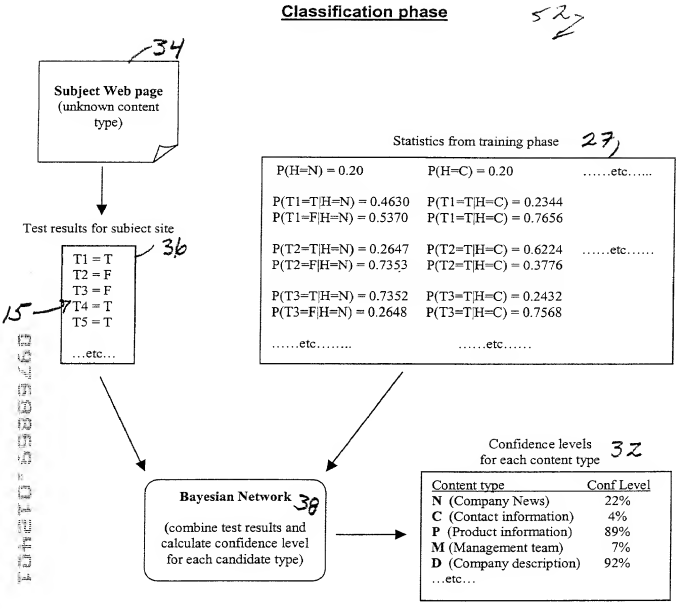
Classification phase
52
34
Subject Web page
(unknown content type)
Statistics from training phase
27
P(H=N) = 0.20
P(H=C) = 0.20
.....etc.....
P(T1=T|H=N) = 0.4630
P(T1=F|H=N) = 0.5370
P(T1=T|H=C) = 0.2344
P(T1=T|H=C) = 0.7656
P(T2=T|H=N) = 0.2647
P(T2=F|H=N) = 0.7353
P(T2=T|H=C) = 0.6224
P(T2=T|H=C) = 0.3776
.....etc.....
P(T3=T|H=N) = 0.7352
P(T3=F|H=N) = 0.2648
P(T3=T|H=C) = 0.2432
P(T3=T|H=C) = 0.7568
.....etc.....
.....etc.....
Test results for subject site
36
T1 = T
T2 = F
T3 = F
T4 = T
T5 = T
...etc...
15
Bayesian Network
38
(combine test results and calculate confidence level for each candidate type)
Confidence levels for each content type
32
Content type
Conf Level
N (Company News) 22%
C (Contact information) 4%
P (Product information) 89%
M (Management team) 7%
D (Company description) 92%
...etc...

Fig. 3

Preferred embodiment

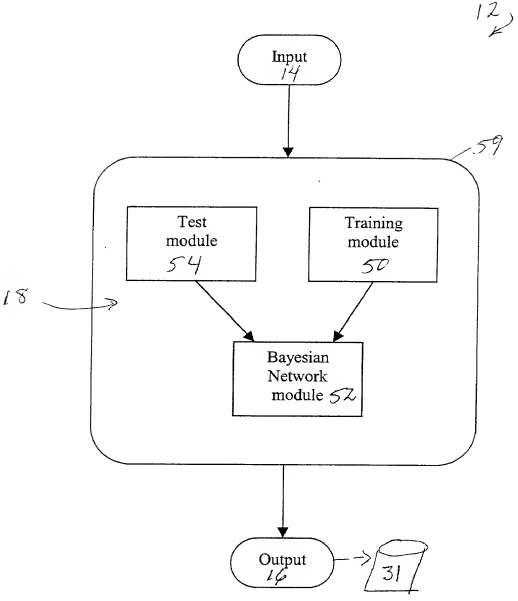

Fig. 4